Revista Chilena de Literatura
ISSN: 0048-7651
rchilite@gmail.com
Universidad de Chile
Chile

Martínez-Gamboa, Ricardo
BIG DATA EN HUMANIDADES DIGITALES: DE LA ESCRITURA DIGITAL A LA "LECTURA DISTANTE"
Revista Chilena de Literatura, núm. 94, 2016, pp. 39-58
Universidad de Chile
Santiago, Chile

REVISTA CHILENA DE LITERATURA
Diciembre 2016, Número 94, 39-58

BIG DATA EN HUMANIDADES DIGITALES: DE LA ESCRITURA DIGITAL A LA "LECTURA DISTANTE"

Ricardo Martínez-Gamboa
Universidad Diego Portales
ricardomartinezg@gmail.com

RESUMEN / *ABSTRACT*

El florecimiento de la escritura digital ha favorecido la irrupción no solo de las humanidades digitales desde el ámbito de la creación literaria, sino que también su análisis por medio de herramientas computacionales. Estas herramientas han permitido establecer tendencias, patrones, regularidades en los textos, a menudo de manera masiva, que han abierto un campo, el de la interpretación mediada por computadores, que ha generado importantes logros a lo largo de las últimas dos décadas, particularmente en lo que respecta a lo que Moretti denomina la "lectura distante". En el presente artículo se exponen algunos de estos logros, a saber: estimaciones cuantitativas temáticas de los textos literarios, estimaciones cuantitativas de contenido de los textos literarios, clasificación de textos literarios según sus propiedades semántico-léxicas, determinación de patrones literarios por género y generación, establecimiento de similitudes entre textos literarios diversos, establecimiento de redes de autores contemporáneos.

Palabras clave: humanidades digitales, lingüística computacional, análisis mediado por computadores, patrones cuantitativos de la literatura.

The development of digital writing has allowed not only the emergence of digital humanities from the field of literary creation, but also its analysis using computer tools. These tools have made it possible to establish trends, patterns, and regularities in the texts, often massively, which have opened a field, that of the interpretation mediated by computers, which has generated significant achievements over the past two decades, particularly in what Moretti has called "distant reading". In this article some of these achievements are presented, namely, quantitative estimates of themes of literary texts, quantitative estimates of content of literary texts, classification of literary texts according to their semantic and lexical properties, the determining of literary patterns by gender and generation, the establishment of similarities between various literary texts, the creation of networks of contemporary authors.

Keywords: digital humanities, computational linguistics, computer mediated analysis, quantitative patterns of literature.

El florecimiento de la escritura digital ha permitido la disponibilidad electrónica de millones de textos a lo largo de las últimas dos décadas. Michel et al. (2011) sostienen que a la fecha (inicios de 2010) se han publicado 129 millones de títulos de libros, de los cuales se encuentran digitalizados, solo en las bases de datos de Google, 15 millones. Estos libros contienen decenas de miles de millones de palabras en todos los idiomas que bien pueden ser indagadas de manera manual, pero, por sobre todo, mediante procedimientos computacionales.

Desde el trabajo seminal de Lancashire (1993) diversos especialistas provenientes de las ciencias de la computación, la lingüística del corpus y la lingüística computacional han intentado llevar a cabo tareas de recuperación de información, análisis, interpretación, detección de patrones (Martínez-Gamboa 2015) y detección de tendencias no solo en los textos digitales de orden general, sino que también, e incluso de manera más intensa, en los textos literarios. Estos esfuerzos suponen que las regularidades en los elementos que son explícitos en las obras literarias (palabras, giros, modismos, estructuras gramaticales, presentación de los personajes, opción por verbos de naturaleza accional o psicológica –Martínez 2012–, etc.) permiten indagar en el contenido y el significado de las obras de manera masiva y robusta (Mitkov 2005; Manning y Schütze 1999). Dicho en simple, cuando se dispone de todas las palabras de un texto en formato digital, es posible aplicar sobre lo textos rutinas de análisis automático que son capaces de detectar patrones que de manera manual sería imposible de determinar u observar. Esta posibilidad abre el campo de los estudios de la literatura hacia un horizonte que tan solo hace unas pocas décadas resultaba insospechado, porque permite establecer regularidades no solo en los textos individuales, sino que en cohortes de escritos literarios de mucho mayor volumen de los que puede revisar un analista particular.

Solo a modo de ejemplo, en un trabajo de esta naturaleza, Soto, Martínez y Sadowsky (2005) hallaron que en los textos de sus respectivas disciplinas escritos por científicos profesionales de las ciencias aplicadas se ocupaban muchos más sustantivos que verbos, mientras que en los escritos de humanidades y ciencias sociales, los profesionales académicos de dichas áreas integraban un número mayor de verbos respecto de sustantivos que sus pares de las ciencias aplicadas y de las ciencias naturales. Este hallazgo, llevado a cabo simplemente analizando decenas de miles de palabras en textos publicados en la base de datos científica nacional Scielo, permite proponer que mientras que los académicos de las ciencias naturales y de las ciencias aplicadas llevan

a cabo procesos de entificación (marcado por la gravitación en el uso de los sustantivos) de los órdenes en los que trabajan, en las ciencias sociales y las humanidades se presta una mayor atención a los procesos (marcado por la gravitación en el uso de los verbos, particularmente aquellos que refieren a acciones).

El principio que opera tras este tipo de indagaciones métricas lingüísticas es que la selección de un rasgo lingüístico explícito (en el anterior caso, la opción entre verbos y sustantivos) da luces para la determinación de ciertas selecciones conceptuales e incluso epistemológicas entre diversos textos. Determinación que sería muy difícil establecer si los textos se hallaran solo en formato impreso, como han mostrado Soto, Sadowsky y Martínez (2014).

Es cierto que la teoría literaria y las disciplinas afines que trabajan con los textos literarios han considerado en una alta estima los procesos de interpretación textual, donde las habilidades de análisis de los críticos, los investigadores, los intérpretes, juegan un papel fundamental. Sin embargo, también es cierto que estos procesos de "close reading" muchas veces son incapaces de tomar en cuenta los grandes números, lo que hoy se denomina el "big data", fundamentalmente porque el manejo de un gran volumen de datos permite observar regularidades y patrones que son muy difíciles de detectar en el microanálisis. Es justamente por ello, que, en un trabajo de posicionamiento de estas nuevas perspectivas, Moretti sostiene que:

> "lo que es una fracción mínima del campo literario es el campo en el que [los teóricos literarios] trabajamos: un canon de 200 novelas, por ejemplo, suena muy grande para el siglo XIX en Gran Bretaña (y es mucho más grande que el actual), pero sigue siendo inferior al uno por ciento de las novelas que se publican en realidad: veinte mil, treinta mil, o más, nadie sabe en realidad –y el "close reading" no ayuda aquí–. Una novela al día, todos los días del año se tardaría en analizar un siglo más o menos (…). [Se propone que] [l]a gran masa [de novelas] sin leer deben ser estudiadas con el fin de ver los patrones de macroevolución en la historia literaria" (Moretti 67).

Hoy los investigadores de la literatura disponen de centenares de miles de textos literarios en formato electrónico a solo un clic de distancia y es posible realizar análisis computacionales con muy pocos recursos computacionales. Es por ello que en lo siguiente se expondrán algunas de las posibles líneas de acción en este campo, a saber: estimaciones cuantitativas temáticas de los textos literarios, estimaciones cuantitativas de contenido de los textos literarios,

clasificación de textos literarios según sus propiedades semántico-léxicas, determinación de patrones literarios por género y generación, establecimiento de similitudes entre textos literarios diversos y establecimiento de redes de autores contemporáneos.

ESTIMACIONES CUANTITATIVAS TEMÁTICAS DE LOS TEXTOS LITERARIOS

Una de las maneras computacionales, y concomitantemente, de humanidades digitales, más sencillas de analizar los textos de modo automatizado y de esta manera aproximarse al contenido semántico-léxico que constituye los textos literarios consiste simplemente en contar las palabras que ocurren con más frecuencia en ellos. Manning y Schütze (21) muestran que una operación de esta naturaleza, por ejemplo, permite determinar el léxico, principalmente de tipo funcional, de una obra como *Las Aventuras de Tom Sawyer*.

Tabla 1

Palabras más frecuentes en *Tom Sawyer*
Manning & Schütze (21)

Word	Freq.	Use
the	3332	determiner (article)
and	2972	conjunction
a	1775	determiner
to	1725	preposition, verbal infinitive marker
of	1440	preposition
was	1161	auxiliary verb
it	1027	(personal/expletive) pronoun
in	906	preposition
that	877	complementizer, demonstrative
he	877	(personal) pronoun
I	783	(personal) pronoun
his	772	(possessive) pronoun
you	686	(personal) pronoun
Tom	679	proper noun
with	642	preposition

Table 1.1 Common words in ***Tom Sawyer.***

Para investigaciones de tipo literario probablemente las métricas de esta naturaleza no resultan suficientemente sofisticadas, como justamente explica la Ley de Zipf, sin embargo, existen diferentes procedimientos mediante los cuales resulta más expedito acceder al contenido temático de los textos, una de ellas es procesar más de un texto a la vez, comparando y contrastando las palabras que en ellos aparecen. Para llevar a cabo este tipo de análisis cuantitativo se suele usar algunos algoritmos específicos, donde probablemente el más eficiente es el método Damerau. El método Damerau consiste en comparar la frecuencia de una palabra en un texto o un conjunto de textos (corpus) con la frecuencia de la misma palabra en otro texto u otro conjunto de textos (corpus). El poder de análisis métrico del método descansa en que la frecuencia relativa resulta ser un índice significativo del contenido semántico de los textos. Luego de aplicar este mecanismo se puede normalizar la aparición de dicha palabra para cada uno de los dos textos o corpora y consiguientemente, establecer la significatividad de su presencia en cada uno.

Martínez (2012) realiza este procedimiento contrastando las palabras en cinco novelas anglosajonas modernas (*Robinson Crusoe*, *Orgullo y Prejuicio*, *La Llamada de la Selva*, *Retrato del Artista Adolescente* y *Matadero Cinco*). Llegando a resultados como los siguientes:

Tabla 2

Palabras más significativas en *Robinson Crusoe*
(en contraste con las restantes cuatro novelas)
[el tamaño relativo de la palabra indica su frecuencia en el texto]

Tabla 3

Palabras más significativas en *Orgullo y Prejuicio*
(en contraste con las restantes cuatro novelas)
[el tamaño relativo de la palabra indica su frecuencia en el texto]

Se puede observar claramente que con este método (Damerau) es posible de manera eficaz, por ejemplo, acceder a los nombres propios de los personajes principales de las novelas, así como a un listado restringido de palabras no funcionales, como sustantivos y verbos que son importantes temáticamente en cada texto. El resultado es un foco en los temas de cada texto (a partir de sus propiedades léxico-semánticas) al que sería muy dificultoso acceder desde la lectura no digital.

Otro aspecto interesante de destacar corresponde a la manera como se ordenan las palabras funcionales en los textos. Martínez (2012) ordena las treinta palabras más frecuentes (prácticamente todas funcionales) en las novelas *Robinson Crusoe*, *Orgullo y Prejuicio*, *La Llamada de la Selva*, *Retrato del Artista Adolescente* y *Matadero Cinco*, como se presenta en la Tabla 4.

Tabla 4

Las treinta palabras más frecuentes en cinco novelas anglosajonas

Robinson Crusoe	Orgullo y Prejuicio	La Llamada de la Selva	Retrato del Artista Adolescente	Matadero 5
the	the	the	the	the
i	to	and	and	and
and	of	of	of	was
to	and	he	a	a
of	her	was	to	to
a	i	to	he	of

my	a	a	his	he
was	in	his	in	in
in	was	in	was	billy
that	she	it	that	it
it	that	buck	had	had
had	it	that	it	that
as	not	with	him	his
for	you	they	you	were
me	he	him	said	on
but	his	had	i	they
with	be	as	with	with
not	as	for	for	so
which	had	on	on	there
he	for	were	from	i
them	with	at	at	said
so	but	but	as	for
this	is	not	is	him
they	have	s	not	from
all	at	by	by	you
or	mr	them	they	all
at	him	from	stephen	as
him	my	out	be	at
be	on	into	but	by
on	s	which	were	she

Se puede observar que el uso de la primera persona en inglés ("I") va descendiendo temporalmente en las cinco novelas (en el caso de *La llamada de la selva* esta palabra es reemplazada por el nombre propio del personaje principal, "Buck"), lo que muestra una tendencia también y permite de manera automática determinar el tipo de narrador en el texto (en primera persona u omnisciente). Por cierto que este aspecto debe ser matizado en parte, puesto que el empleo de este tipo de métodos solo apunta a la adopción de modos de razonamiento inductivo y estadístico en los estudios.

ESTIMACIONES CUANTITATIVAS DE CONTENIDO DE LOS TEXTOS LITERARIOS

Un segundo modo de aproximación al contenido de los textos literarios por medio de procesos de lingüística computacional y análisis de "lectura distante" consiste en agrupar las palabras de acuerdo con sus características semánticas propiamente tales. El estándar como proceso para llevar a cabo esta tarea es el etiquetado semántico que se puede realizar con el *software* LIWC.

Se trata de un *software* que:

> "Se compone de cerca de 4.500 palabras y palabras derivadas. Cada raíz de palabra o palabra define una o más categorías. Por ejemplo, la palabra *cried* forma parte de cinco categorías de palabras: *tristeza*, *emoción negativa*, *afecto*, *verbo* y el *verbo en tiempo pasado*" (Pennebacker et al. 2007).

Estas palabras forman 64 categorías gramaticales y semánticas y que permiten, subsecuentemente, realizar un análisis robusto de contenidos de los textos y en 2011 se ha elaborado por el equipo desarrollador un sistema para el español.

Con estos medios, Martínez (2012) ha analizado las cinco novelas anglosajonas mencionadas anteriormente, llegando a las siguientes distribuciones:

Tabla 5

Contenidos de cinco novelas anglosajonas – Frecuencia cada cien palabras (en gris claro la novela con menor presencia del contenido y en gris oscuro la novela con más presencia del contenido)

	Robinson Crusoe	Orgullo y Prejuicio	La Llamada de la Selva	Retrato del Artista Adolescente	Matadero 5
Conectores	64,84	61,75	55,6	54,83	55,84
Pronombres	17,36	16,98	11,53	13,27	12,81
Pronombres Impersonales	5,72	4,85	3,83	3,96	4,95
Artículos	6,75	5,43	9,68	9,46	8,75
Verbos	13,33	14,9	9,49	11,23	14,61
Verbos Auxiliares	7,29	9,84	5,62	6,24	9,13

Groserías	0,01	0	0,04	0,2	0,15
Sociedad	7,06	16,14	10,23	12,93	11,55
Familia	0,12	1,13	0,17	0,52	0,41
Amistad	0,11	0,3	0,17	0,25	0,14
Humanos	0,42	1,75	0,66	0,86	0,98
Afectividad	4,14	6,63	4,56	5,05	3,89
Emociones Positivas	2,39	4,58	1,98	2,56	1,9
Emociones Negativas	1,72	2,02	2,56	2,44	1,98
Ansiedad	0,43	0,49	0,49	0,46	0,22
Procesos Cognitivos	17,55	17,95	14,05	13,51	12,57
Procesos Perceptuales	1,86	2,11	2,88	4,8	3,19
Procesos Biologicos	1,58	0,79	3,54	3,02	2,56
Sexualidad	0,04	0,11	0,19	0,23	0,19
Ingestion	0,47	0,16	0,45	0,4	0,52
Movimiento	2,46	1,79	2,8	2	2,44
Trabajo	0,69	0,75	0,78	0,9	1,34
Logro	1,2	1,4	1,46	0,94	0,87
Esparcimiento	0,35	0,61	0,89	0,71	1,18
Hogar	0,2	0,56	0,17	0,51	0,81
Dinero	0,38	0,41	1,28	0,26	0,46
Religion	0,45	0,21	0,18	2,04	0,67
Muerte	0,29	0,07	0,34	0,33	0,86

Como se puede observar en la tabla, es posible determinar los núcleos temáticos de cada texto en general, donde, por ejemplo, la muerte es un tema en *Matadero 5*, pero no (en contraste) en *Orgullo y Prejuicio*, mientras que la sexualidad está más presente en *Retrato del Artista Adolescente* que en las restantes cuatro novelas. Lancashire (1993) ha mostrado que procesos similares pueden ser llevados a cabo al interior de las obras literarias, por ejemplo, comparando el comportamiento del léxico de contenido entre dos capítulos diferentes de la obra, como ha hecho dicho autor respecto de *El Cuento de la Sirvienta* de Margaret Atwood, donde descubrió que era posible distinguir

entre capítulos diurnos y nocturnos en el texto que contrastaban tanto en sus contenidos como en su temple de ánimo. Asimismo, los investigadores que realizan análisis de contenido de esta manera pueden elegir la manera como segmentan el texto (Mikheev 2003), por lo que el método resulta aplicable también para el estudio de otros tipos de texto aparte de las novelas, como los cuentos o los poemas.

CLASIFICACIÓN DE TEXTOS LITERARIOS SEGÚN SUS PROPIEDADES SEMÁNTICO-LÉXICAS

Una manera más elaborada de procesar los textos literarios en cuanto a las palabras que ocupan consiste en determinar tendencias o patrones generales en textos que pertenecen a un mismo periodo literario o a un mismo género. Para ello diversos autores (Biber 1991; Martínez-Gamboa 2015) han propuesto el uso de procedimientos de análisis factorial de los rasgos (léxicos y gramaticales) que se presentan en los textos. Someramente, el análisis factorial determina conjuntos de palabras que pertenecen a las mismas categorías (como el uso de la primera persona o ítems léxicos vinculados al dominio semántico de la "familia"), y, posteriormente agrupa dichos rasgos en coocurrencias (por ejemplo, que el uso de determinantes coocurre en un texto con el uso de la tercera persona gramatical, entre otras muchas posibilidades):

> "La utilidad más relevante de este tipo de análisis es el hecho de que provee a los investigadores de una representación de la estructura subyacente de los datos estadísticamente fundamentada. Esta estructura viene dada por un número lo más pequeño posible de dimensiones de covariación de factores" (Martínez 154).

De este modo, por ejemplo, en el análisis de treinta novelas chilenas que cubren un periodo que va de 1862 (*Martín Rivas*) a 2011 (*Ruido*), Martínez (2012) ha logrado establecer que dichas novelas se agrupan en tres grandes grupos (dimensiones) en cuanto a sus rasgos:

- Dimensión 1: Narración-Descripción
- Dimensión 2: Cognición-Emoción
- Dimensión 3: Social-Individual

Las novelas, de acuerdo con el procedimiento, obtienen un puntaje (que se establece en una polaridad positiva o negativa). Por ejemplo, *El niño que enloqueció de amor* obtiene un puntaje de 35,06 en la Dimensión 1 (Narración-Descripción), siendo la novela más narrativa de las analizadas; mientras que *Ruido* obtiene un puntaje de -32,61, siendo la novela menos narrativa y más descriptiva del conjunto. La Tabla 6 (Martínez-Gamboa 247) muestra esta distribución para la Dimensión 1. Resulta necesario indicar que el método, usado originalmente por Biber para clasificar variedades textuales se emplea acá para categorizar individuos textuales. Eso vuelve, por cierto, decisivas cuestiones metodológicas como la selección de la muestra.

Tabla 6

Valores en la Dimensión 1 (Narración-Descripción)

El niño que enloqueció de amor	35,06
Papelucho detective	23,34
Papelucho	22,95
Lumpérica	22,22
Palomita Blanca	19,99
Mala onda	16,41
Papelucho y el marciano	13,53
Formas de volver a casa	11,18
Papelucho perdido	9,35
Patas de perro	9,12
Papelucho en vacaciones	7,28
Papelucho historiador	6,50
Martín Rivas	5,57
Los detectives salvajes	3,44
El lugar sin límites	2,45
La ultima niebla	1,04
Eloy	0,17
Estrellas Muertas	-1,97
Hijo de Ladrón	-2,55
La amortajada	-7,93

El socio	-8,09
El loco Estero	-9,37
Bonsái	-10,65
Música Marciana	-16,78
2666	-17,81
La casa de los espíritus	-19,38
Condell	-25,89
El último grumete de la Baquedano	-25,94
Alhué	-30,62
Ruido	-32,61

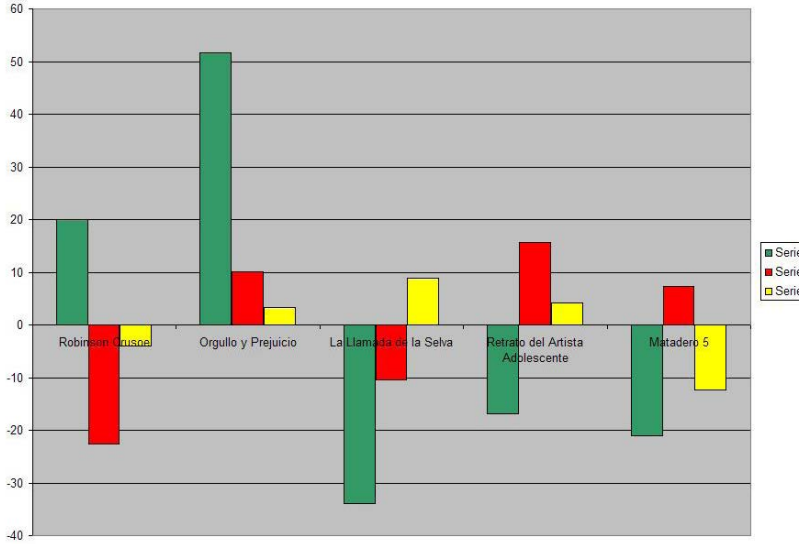
La utilidad de estas clasificaciones es que permite ordenar las novelas (y los textos literarios en general) en conglomerados mayores de manera robusta, como se indicará a continuación.

DETERMINACIÓN DE PATRONES LITERARIOS POR GÉNERO Y GENERACIÓN

Una vez que se dispone de clasificaciones cuantitativo cualitativas de la obras literarias de acuerdo con sus rasgos gramático-semántico-léxicos, es posible establecer conjuntos de obras o patrones generales. Por ejemplo, en el caso de las cinco novelas anglosajonas revisadas en Martínez (2012), es posible observar cómo ellas se organizan en cuanto a las Dimensiones de coviariación:

Tabla 7

Características generales de cinco novelas anglosajonas



Donde la Serie 1 (Dimensión 1) señala el eje Emociones Positivas - Emociones Negativas[1], la Serie 2 (Dimensión 2) señala el eje Humanización - Deshumanización[2] y la Serie 3 señala el eje Victoria Pírrica - Triunfo Moral[3], de acuerdo con el análisis de Martínez (2012).

Se puede observar que las novelas, en la medida en que transcurre el tiempo, van haciéndose cada vez más negativas (Serie 1).

Un resultado similar, que muestra una tendencia hacia la desensibilización, es el obtenido por Martínez-Gamboa (2015):

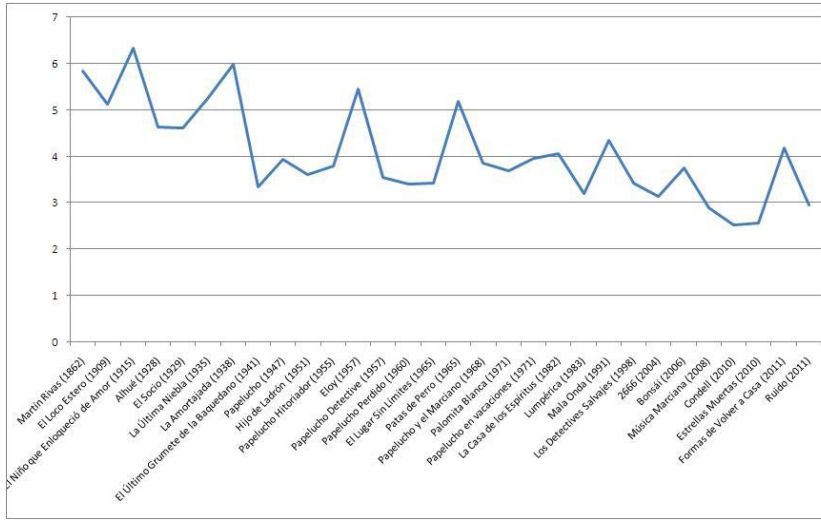
[1] Dimensión 1: Emociones Positivas - Emociones Negativas. En ella son extremadamente positivos aspectos de los textos como los procesos cognitivos, las emociones positivas, la afectividad y la familia; y extremadamente negativos aspectos de los textos como las groserías, el enojo, las emociones negativas y los procesos biológicos.

[2] Dimensión 2: Humanización - Deshumanización. En ella son extremadamente positivos aspectos como escuchar, hogar, sociedad, asentimiento y amistad; y extremadamente negativos aspectos como números, preposiciones y conjunciones.

[3] Dimensión 3: Victoria Pírrica - Triunfo Moral. En un extremo se encuentran aspectos como logro, ansiedad, tristeza y emociones negativas; y en el otro, aspectos como hogar, muerte, trabajo, pasado y verbos.

Tabla 8

Desensibilización de las novelas chilenas en el tiempo

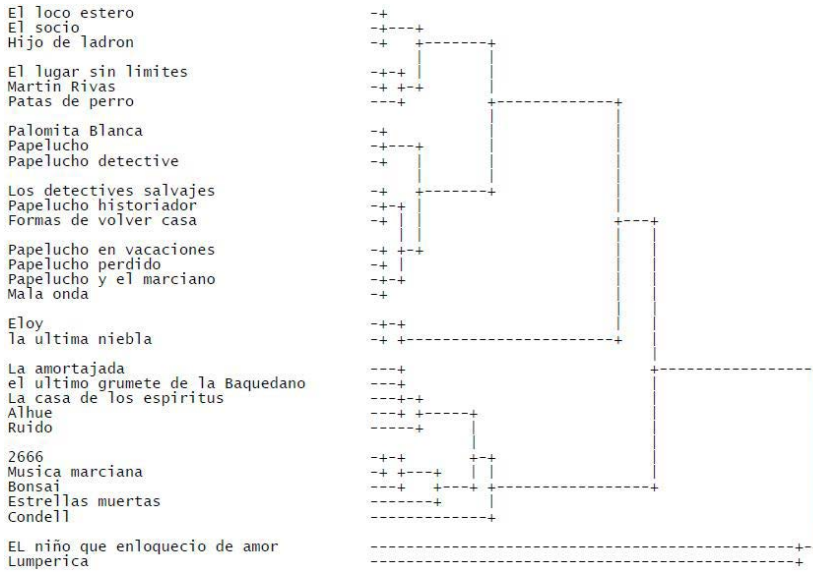


En este caso se ha procedido a contabilizar la frecuencia de las palabras afectivas (como acariciar, aburrir, fracaso, etc.) y se han ordenado cronológicamente las treinta novelas. La tendencia es hacia el descenso de dicho tipo de palabras (más frecuentes en *El niño que enloqueció de amor*, y menos frecuente en las novelas de la presente década). Por cierto que este resultado es exploratorio, por lo que se requiere de un corpus aún más masivo para consolidarlo.

Otra manera de determinar tendencias es agrupar las novelas de acuerdo con sus rasgos en lo que se llaman espacios vectoriales. Por ejemplo, en el caso del estudio de Martínez-Gamboa (2015) donde se establecieron las tres dimensiones anteriormente mencionadas es posible agrupar a las novelas con un procedimiento que se denomina de clusterización. El resultado es el siguiente (se presenta como un *dendrograma* que agrupa las novelas en conjuntos pares en que la tabla indica, con un signo "+" el nivel de cercanía de parejas de novelas o grupos de ellas):

Tabla 9

Análisis de *cluster* de treinta novelas chilenas



Como se señala en Martínez-Gamboa (249), en el dendrograma es posible observar dos conjuntos de agrupaciones bastante claros. El primero, desde *El loco estero* hasta *Mala onda* y el segundo desde *La amortajada* hasta *Condell*. Quedan fuera de estos dos grupos dos novelas consideradas unánimemente por la crítica literaria como anómalas: *El niño que enloqueció de amor* (cf. Vásquez 1963) y *Lumpérica* (cf. Donoso 2009); la primera por su exceso de emocionalidad y la segunda por su explícita y expresa prosa rupturista. Dentro de estos dos conjuntos mayores se observa que el primero corresponde casi íntegramente a obras del siglo XX, mientras que en el segundo se aglomeran las obras del bolañismo y el posbolañismo (*2666*, *Ruido*, *Música marciana*, *Bonsái*, etc.). Del mismo modo, es posible advertir que en un conjunto de relaciones dendrográmicas bastante estrechas (lo que es marcado en el diagrama por la cercanía de las marcas "+" al lado izquierdo del mismo) entre *El loco Estero*, *El socio* e *Hijo de ladrón*, obras todas de la primera mitad del siglo XX, y más abajo, entre *El lugar sin límites*, *Martín Rivas* y *Patas de perro*, todas obras, ellas y las tres anteriores, de la "masculinidad". Los siguientes tres grupos (esto es, desde *Palomita Blanca* hasta *Mala onda*) constituyen

un grupo que se podría denominar de novelas de formación (*Bildungsroman*, cf. Arango 2009), nucleadas en torno a los libros de Papelucho.

ESTABLECIMIENTO DE SIMILITUDES ENTRE TEXTOS LITERARIOS DIVERSOS

Finalmente, respecto de las obras narrativas, es posible establecer similitudes entre ellas, toda vez que al distribuirlas en espacios vectoriales es posible determinar la distancia que guarda cada dupla o par de novelas. Este procedimiento se lleva a cabo por medio de la siguiente fórmula (los valores son los valores numéricos de cada Dimensión):

$$\cos \alpha = \frac{u_1 . v_1 + u_2 . v_2 + u_3 . v_3}{\sqrt{u_1^2 + u_2^2 + u_3^2} \cdot \sqrt{v_1^2 + v_2^2 + v_3^2}}$$

Donde el coseno alfa adquiere valores que van de -1 para la absoluta disimilitud a +1 para la absoluta similitud.

Estos son los valores para las treinta novelas analizadas en Martínez-Gamboa (2015):

Tabla 10
Distancias coseno alfa entre las novelas

	2666	Alhue	Bonsai	Condell	El loco estero	El lugar sin límites	EL niño que enloquecio de amor	El socio	El ultimo grumete de la Baguedano	Eloy	Estrellas muertas	Formas de volver casa	Hijo de ladron	La amortajada	la ultima niebla	La casa de los espiritus	Los detectives salvajes	Lumperica	Mala onda	Martin Rivas	Musica marciana	Palomita Blanca	Papelucho detective	Papelucho en vacaciones	Papelucho historiador	Papelucho perdido	Papelucho y el marciano	Papelucho	Patas de perro	Ruido
2666	1,00	0,75	0,91	0,91	0,47	-0,76	-1,00	0,48	0,91	-0,52	0,61	-0,28	0,63	-0,26	-0,53	0,75	0,18	-0,33	-0,68	-0,68	1,00	-0,64	-0,72	-0,72	-0,10	-0,68	-0,90	-0,75	-0,88	0,96
Alhue	0,75	1,00	0,54	0,58	0,91	-0,25	-0,80	0,88	0,91	0,16	-0,05	-0,84	0,60	0,42	0,09	0,99	-0,51	-0,53	-0,97	-0,36	0,80	-0,95	-1,00	-0,99	-0,73	-0,98	-0,96	-0,98	-0,35	0,90
Bonsai	0,91	0,54	1,00	1,00	0,16	-0,95	-0,88	0,14	0,84	-0,58	0,64	-0,05	0,28	-0,35	-0,50	0,59	0,42	-0,50	-0,55	-0,91	0,88	-0,34	-0,52	-0,56	0,08	-0,55	-0,69	-0,48	-0,94	0,79
Condell	0,91	0,58	1,00	1,00	0,19	-0,94	-0,88	0,16	0,87	-0,52	0,59	-0,10	0,25	-0,28	-0,43	0,63	0,37	-0,56	-0,59	-0,93	0,88	-0,36	-0,56	-0,60	0,02	-0,59	-0,70	-0,50	-0,92	0,80
El loco estero	0,47	0,91	0,16	0,19	1,00	0,16	-0,55	0,99	0,65	0,39	-0,29	-0,94	0,67	0,59	0,25	0,87	-0,77	-0,28	-0,85	0,06	0,55	-0,98	-0,91	-0,86	-0,83	-0,85	-0,81	-0,94	0,00	0,71
El lugar sin límites	-0,76	-0,25	-0,95	-0,94	0,16	1,00	0,71	0,18	-0,63	0,71	-0,75	-0,26	-0,08	0,54	0,59	-0,31	-0,67	0,40	0,27	0,92	-0,71	0,03	0,23	0,28	-0,35	0,27	0,43	0,18	0,95	-0,57
EL niño que enloquecio de amor	-1,00	-0,80	-0,88	-0,88	-0,55	0,71	1,00	-0,55	-0,93	0,45	-0,55	0,36	-0,66	0,19	0,48	-0,80	-0,10	0,35	0,73	0,64	-1,00	0,71	0,77	0,77	0,17	0,73	0,93	0,80	0,83	-0,98
El socio	0,48	0,88	0,14	0,16	0,99	0,18	-0,55	1,00	0,61	0,33	-0,23	-0,89	0,74	0,52	0,17	0,83	-0,74	-0,18	-0,80	0,12	0,55	-0,98	-0,88	-0,82	-0,78	-0,80	-0,81	-0,94	-0,01	0,70
El ultimo grumete de la Baguedano	0,91	0,91	0,84	0,87	0,65	-0,63	-0,93	0,61	1,00	-0,14	0,24	-0,58	0,45	0,14	-0,13	0,93	-0,13	-0,65	-0,91	-0,71	0,93	-0,76	-0,90	-0,92	-0,46	-0,91	-0,93	-0,84	-0,67	0,95
Eloy	-0,52	0,16	-0,58	-0,52	0,39	0,71	0,45	0,33	-0,14	1,00	-0,99	-0,65	-0,33	0,96	0,97	0,17	-0,87	-0,36	-0,27	0,39	-0,46	-0,20	-0,21	-0,22	-0,79	-0,27	0,13	-0,08	0,81	-0,28
Estrellas muertas	0,61	-0,05	0,64	0,59	-0,29	-0,75	-0,55	-0,23	0,24	-0,99	1,00	0,57	0,40	-0,92	-0,98	-0,06	0,82	0,31	0,17	-0,44	0,55	0,10	0,10	0,12	0,72	0,16	-0,24	-0,03	-0,86	0,38
Formas de volver casa	-0,28	-0,84	-0,05	-0,10	-0,94	-0,26	0,36	-0,89	-0,58	-0,65	0,57	1,00	-0,36	-0,82	-0,56	-0,83	0,89	0,50	0,86	-0,03	-0,35	0,87	0,87	0,85	0,97	0,86	0,66	0,81	-0,20	-0,54
Hijo de ladron	0,63	0,60	0,28	0,25	0,67	-0,08	-0,66	0,74	0,45	-0,33	0,40	-0,36	1,00	-0,18	-0,51	0,51	-0,18	0,34	-0,41	0,13	0,67	-0,75	-0,56	-0,46	-0,15	-0,41	-0,73	-0,76	-0,38	0,71
La amortajada	-0,26	0,42	-0,35	-0,28	0,59	0,54	0,19	0,52	0,14	0,96	-0,92	-0,82	-0,18	1,00	0,93	0,44	-0,91	-0,53	-0,53	0,21	-0,19	-0,43	-0,47	-0,49	-0,93	-0,53	-0,14	-0,33	0,63	0,00
la ultima niebla	-0,53	0,09	-0,50	-0,43	0,25	0,59	0,48	0,17	-0,13	0,97	-0,98	-0,56	-0,51	0,93	1,00	0,13	-0,75	-0,49	-0,24	0,23	-0,48	-0,07	-0,14	-0,19	-0,73	-0,24	0,21	0,03	0,75	-0,33
La casa de los espiritus	0,75	0,99	0,59	0,63	0,87	-0,31	-0,80	0,83	0,93	0,17	-0,06	-0,83	0,51	0,44	0,13	1,00	-0,48	-0,63	-0,99	-0,45	0,80	-0,91	-0,99	-1,00	-0,73	-0,99	-0,94	-0,94	-0,38	0,90
Los detectives salvajes	0,18	-0,51	0,42	0,37	-0,77	-0,67	-0,10	-0,74	-0,13	-0,87	0,82	0,89	-0,18	-0,91	-0,75	-0,48	1,00	0,23	0,53	-0,44	0,10	0,62	0,54	0,51	0,92	0,53	0,27	0,50	-0,62	-0,11
Lumperica	-0,33	-0,53	-0,50	-0,56	-0,28	0,40	0,35	-0,18	-0,65	-0,36	0,31	0,50	0,34	-0,53	-0,49	-0,63	0,23	1,00	0,70	0,72	-0,34	0,29	0,56	0,66	0,57	0,70	0,39	0,33	0,20	-0,39
Mala onda	-0,68	-0,97	-0,55	-0,59	-0,85	0,27	0,73	-0,80	-0,91	-0,27	0,17	0,86	-0,41	-0,53	-0,24	-0,99	0,53	0,70	1,00	0,46	-0,73	0,88	0,98	1,00	0,78	1,00	0,89	0,90	0,30	-0,84
Martin Rivas	-0,68	-0,36	-0,91	-0,93	0,06	0,92	0,64	0,12	-0,71	0,39	-0,44	-0,03	0,13	0,21	0,23	-0,45	-0,44	0,72	0,46	1,00	-0,64	0,09	0,36	0,45	-0,06	0,45	0,44	0,22	0,79	-0,55
Musica marciana	1,00	0,80	0,88	0,88	0,55	-0,71	-1,00	0,55	0,93	-0,46	0,55	-0,35	0,67	-0,19	-0,48	0,80	0,10	-0,34	-0,73	-0,64	1,00	-0,70	-0,77	-0,76	-0,17	-0,73	-0,93	-0,80	-0,84	0,98
Palomita Blanca	-0,64	-0,95	-0,34	-0,36	-0,98	0,03	0,71	-0,98	-0,76	-0,20	0,10	0,87	-0,75	-0,43	-0,07	-0,91	0,62	0,29	0,88	0,09	-0,70	1,00	0,95	0,90	0,73	0,88	0,91	0,99	0,20	-0,83
Papelucho detective	-0,72	-1,00	-0,52	-0,56	-0,91	0,23	0,77	-0,88	-0,90	-0,21	0,10	0,87	-0,56	-0,47	-0,14	-0,99	0,54	0,56	0,98	0,36	-0,77	0,95	1,00	0,99	0,76	0,98	0,94	0,97	0,32	-0,88
Papelucho en vacaciones	-0,72	-0,99	-0,56	-0,60	-0,86	0,28	0,77	-0,82	-0,92	-0,22	0,12	0,85	-0,46	-0,49	-0,19	-1,00	0,51	0,66	1,00	0,45	-0,76	0,90	0,99	1,00	0,76	1,00	0,92	0,93	0,33	-0,87
Papelucho historiador	-0,10	-0,73	0,08	0,02	-0,83	-0,35	0,17	-0,78	-0,46	-0,79	0,72	0,97	-0,15	-0,93	-0,73	-0,73	0,92	0,57	0,78	-0,06	-0,17	0,73	0,76	0,76	1,00	0,78	0,49	0,66	-0,36	-0,36
Papelucho perdido	-0,68	-0,98	-0,55	-0,59	-0,85	0,27	0,73	-0,80	-0,91	-0,27	0,16	0,86	-0,41	-0,53	-0,24	-0,99	0,53	0,70	1,00	0,45	-0,73	0,88	0,98	1,00	0,78	1,00	0,89	0,90	0,30	-0,84
Papelucho y el marciano	-0,90	-0,96	-0,69	-0,70	-0,81	0,43	0,93	-0,81	-0,93	0,13	-0,24	0,66	-0,73	-0,14	0,21	-0,94	0,27	0,39	0,89	0,44	-0,93	0,91	0,94	0,92	0,49	0,89	1,00	0,97	0,58	-0,99
Papelucho	-0,75	-0,98	-0,48	-0,50	-0,94	0,18	0,80	-0,94	-0,84	-0,08	-0,03	0,81	-0,76	-0,33	0,03	-0,94	0,50	0,33	0,90	0,22	-0,80	0,99	0,97	0,93	0,66	0,90	0,97	1,00	0,35	-0,91
Patas de perro	-0,88	-0,35	-0,94	-0,92	0,00	0,95	0,83	-0,01	-0,67	0,81	-0,86	-0,20	-0,38	0,63	0,75	-0,38	-0,62	0,20	0,30	0,79	-0,84	0,20	0,32	0,33	-0,36	0,30	0,58	0,35	1,00	-0,70
Ruido	0,96	0,90	0,79	0,80	0,71	-0,57	-0,98	0,70	0,95	-0,28	0,38	-0,54	0,71	0,00	-0,33	0,90	-0,11	-0,39	-0,84	-0,55	0,98	-0,83	-0,88	-0,87	-0,36	-0,84	-0,99	-0,91	-0,70	1,00

ESTABLECIMIENTO DE REDES DE AUTORES CONTEMPORÁNEOS

Un último procedimiento que merece ser presentado es el que respecta ya no al análisis textual, sino que al análisis del campo literario, por ejemplo, lo referido a las generaciones o redes de escritoras y escritores. En un trabajo en preparación (Martínez-Gamboa & Gaínza, en preparación) se ha llevado a cabo dicho análisis. Se han seleccionado los veintiún escritores chilenos con más visitas de páginas web de Chile (.cl), entre los que han publicado los libros de relatos y/o novelas cortas en el periodo de cinco años que media entre 2009 y 2014. Se han añadido las fechas de nacimiento de los escritores para dejar claro el orden que emerge.

El gráfico muestra las relaciones entre ellos, es decir, las entradas comunes en los sitios web. Las líneas que unen dos escritores indican cuán a menudo aparecen juntos en los sitios web que los mencionan, y el grosor de la línea que conecta los representa el "peso" de estas menciones conjuntas. Esta cifra se ha analizado con el programa Pajek que permite el análisis de las "redes sociales" y se ha ordenado la figura con el modelo de la energía Kamada-Kawai que imprime un "orden" a las relaciones entre los nodos, en este caso, los escritores. Las figuras se forman en "constelaciones", es decir, grupos de escritores que mantienen una cercanía entre ellos.

Como puede observarse, en esta versión del mapa surgen tres conjuntos de escritores fuertemente interrelacionados.

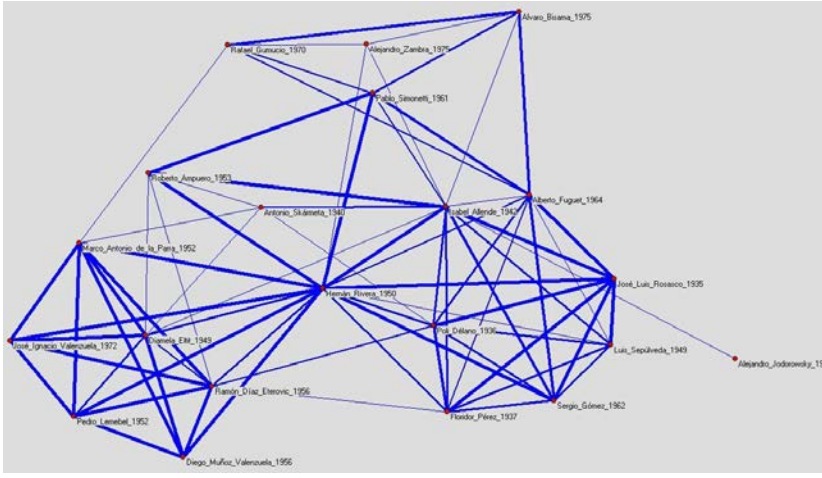
Un primer grupo está formado por los escritores nacidos en los años veinte, treinta y cuarenta, abajo a la derecha. Estos escritores se hallan nucleados en torno a la figura de Isabel Allende.

Un segundo grupo son los escritores de los años cincuenta que están nucleados en torno a la figura de Damiela Eltit.

Por último, se observa que surge como un tercer grupo o constelación de escritores de los años sesenta y setenta, sin núcleo claro y más relacionado con el primer grupo que en el segundo.

Cabe señalar que en el centro de la figura se forma una "constelación" de lo que se podría denominar "escritores superventa".

Gráfico 1. El mapa de la literatura chilena actual



CONCLUSIÓN

Tomando en consideración los métodos y tipos de análisis presentados es notorio que las técnicas de lingüística computacional y del corpus computacional permiten acceder, cuando se aplica a los textos literarios y a los movimientos y escenas literarias, a aspectos de los textos y campos que sería sumamente engorroso trabajar solo mediante el "close reading". Estos procesos de lectura distante iluminan dimensiones de la literatura que estaban vedados a los analistas antes del advenimiento de las humanidades digitales y hoy son mecanismos que resultan de aplicación sencilla y resultados prometedores en la mayoría de los casos. No se trata solamente de que el "close reading" vuelva engorroso cierto análisis, sino que lo que se propone en la presente investigación es un paradigma diferente del análisis de textos literarios, en el que los aspectos cuantitativos y el método inductivo desempeñan un papel crucial. Esto implica no solo el empleo de herramientas computacionales, sino también la adopción de diseños metodológicos apropiados a este tipo de análisis.

BIBLIOGRAFÍA

Arango, Selen. "La novela de formación y sus relaciones con la pedagogía y los estudios literarios", *Folios* 30 (2009): 127-146.

Biber, Douglas. *Variation across speech and writing*. Cambridge University Press, 1991.

Donoso, Jaime. "Práctica de la Avanzada: *Lumpérica* y la figuración de la escritura como fin de la representación burguesa de la literatura y el arte". *Diamela Eltit: redes locales, redes globales*, Madrid: Iberoamericana (2009): 239-260.

Lancashire, Ian. "Computer-assisted critical analysis: a case study of Margaret Atwood's Handmaid's Tale." *The digital word*. MIT Press (1993): 293-318.

Manning, Christopher D., and Hinrich Schütze. *Foundations of statistical natural language processing.* Cambridge: MIT Press, 1999.

Martínez, Ricardo. "Can you read my mind?: la historia de la mente en cinco novelas anglosajonas modernas." *Where is my mind?* Cuarto Propio (2012): 135-170.

Martínez-Gamboa, Ricardo. "Patrones cuantitativos en novelas chilenas de los siglos XIX a XXI." *Onomázein* 2.32 (2015): 239-253.

Martínez-Gamboa, Ricardo, y Carolina Gaínza. "El mapa de la literatura chilena actual". En preparación.

Michel, Jean-Baptiste et al. "Quantitative analysis of culture using millions of digitized books". *Science* 331.6014 (2011): 176-182.

Mikheev, Andrei. "Text segmentation". *The Oxford handbook of computational linguistics* (2003): 201-218.

Mitkov, Ruslan. *The Oxford handbook of computational linguistics*. Oxford University Press, 2005.

Moretti, Franco. *Graphs, Maps, Trees: Abstract Models for a Literary Theory*. Verso, 2005.

Pennebaker, James W., Cindy K. Chung, Molly Ireland, Amy Gonzales y Roger J. Booth. *The development and psychometric properties of LIWC2007*. Austin, Texas: LIWC.net, 2007.

Soto Vergara, Guillermo, Ricardo Martínez y Scott Sadowsky. "Verbos y sustantivos en textos científicos. Análisis de variación en un corpus de textos de ciencias aplicadas, naturales, sociales y humanidades." *Philologia hispalensis* 19 (2005): 169-187.

Soto Vergara, Guillermo, Scott Sadowsky y Ricardo Martínez Gamboa. "El le invariable en el español escrito de Chile". *Literatura y lingüística* 29 (2014): 214-225.

Vásquez, Ángel. "Los tres planos de la creacion artística de Eduardo Barrios". *Revista Iberoamericana XXIX* 55 (1963): 125-37.